Sahil Maisuria
Professor Wallisch
Big Data Analysis Project
May 19th, 2021

Data Handling and Dimension Reduction

Upon prior inspection, it was revealed that about 21.67% of the data was missing. The per pupil spending and average class size data on charter schools were systematically missing and imputation would not be the correct choice here since replacing missing values with either the mean or median would severely limit variability as charter schools are very different from public schools. Other randomly missing data points were dropped row-wise before computing correlations, linear regression, and dimensionality reduction using Principal Component Analysis. Otherwise, data were dropped element wise to conduct independent sample t-tests. PCA was used to reduce the features from the school climate variables and achievement indicators into a smaller set that still contained most of the variance explained by the original predictors. The Kaiser criterion (eigenvalues > 1) was used to select independent factors.

Exploratory Data Analysis

The data on the number of applications, acceptances, math scores, and several of the demographic indicators were skewed to right. A log linearisation was appropriate for all continuous variables except for the number of applications and admissions since even after conducting the transformation, the data was still not close to being approximately normal.

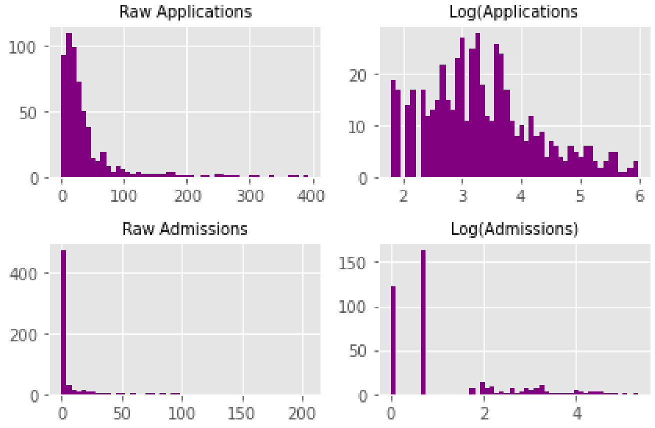


Fig 1: Data transformation

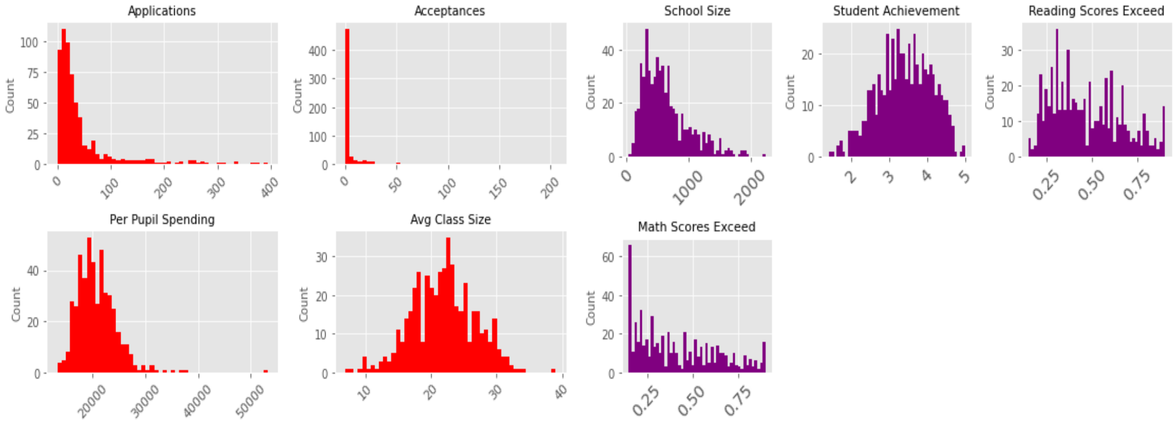


Fig 2 & 3: Distribution of some important predictors

Question 1 and Question 2:

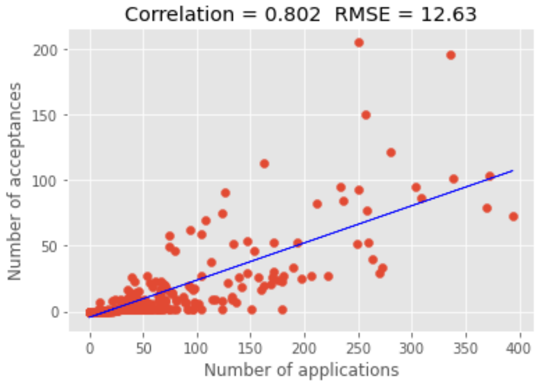


Fig 4: # of Applications

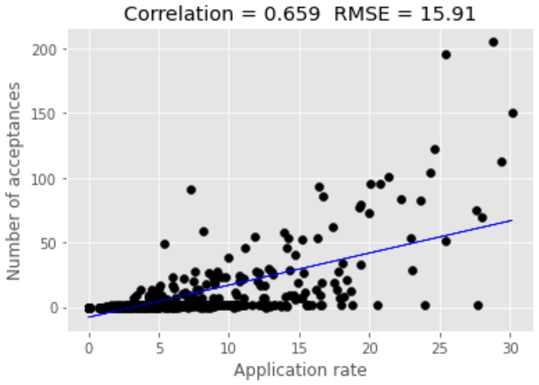


Fig 5: Application rate

The spearman correlation between the number of applications and admissions to HSPHS was revealed to be 0.802 which implies that there is a strong and positive linear relationship between the number of applications, and the number of acceptances. Meaning, that all else equal, if a given school increases its number of applications to HSPS, then there is an increase in the number of acceptances, though this also depends on several other factors which will be explained further on. To compare whether number of applications or the application rate is a better predictor of admission to HSPS, we must first compute the application rate as: $application\ rate\ \left(\frac{\#\ of\ applications}{school\ size}\right) * 100$. Then, we fit a linear regression line, and see which measure has more variance explained, and a lower RMSE. Based on this, it was revealed that, about 64.28% of the variance was explained by the model with raw number of applications in contrast to 43.4% of the variance explained by the model with application rate. Both predictors were statistically significant as their p-values were lower than $\alpha = 0.05$ at a 95% significance level. Thus, it seems that the raw number of applications is a better predictor of admission to HSPS since it explains more variance, and using it in our model produces a lower RMSE score of 12.63.

Question 3:

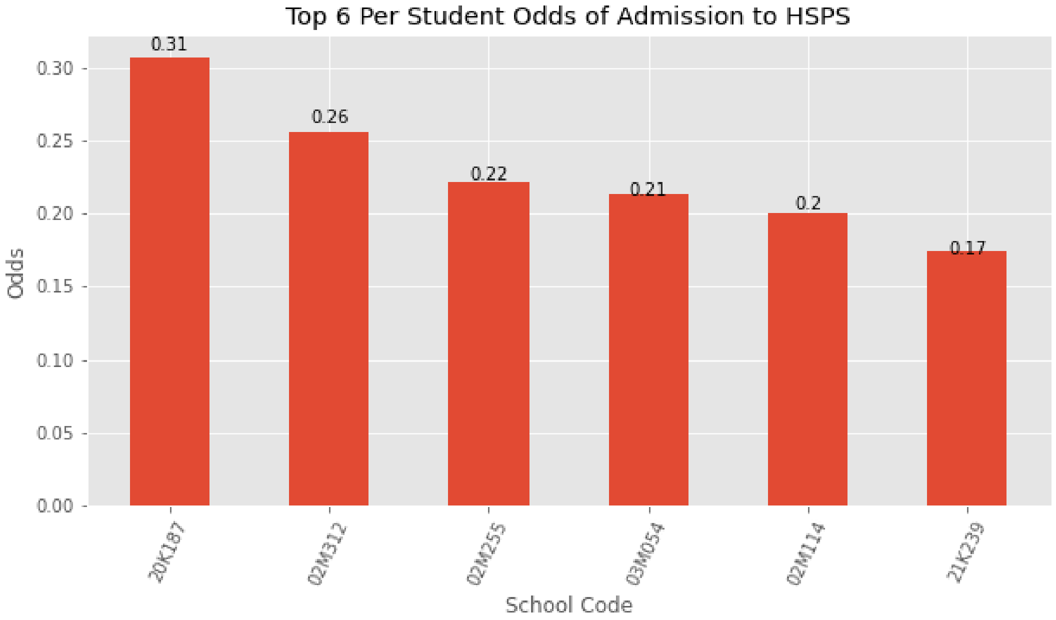


Fig 6: Top 6 schools with the best per student odds of admission to HSPS

To compute which school has the best per student odds of being accepted to HSPS, we compute it as: $Odds = \frac{\left(\frac{Acceptances}{School\,Size}\right)}{\left(1-\frac{Acceptances}{School\,Size}\right)}$. Based on this, the school with the best per student odds of sending someone to HSPS is The Christa Mcauliffe School\I.S. 187. It had 251 applications, 205 acceptances, and a school size of 873, with per student odds of 0.307:1.

Question 4:

To test if there is a relationship between the school climate variables, and how student perform on objective measures of achievement, we first have to reduce the dimensions of each set features using a PCA, and find the uncorrelated features that explain most of the variance. In order to do that we first extract all the school climate variables, and achievement indicators in a data frame, and remove any missing values to preserve their lengths. Next, run the PCA, analyse the scree plot and select independent factors based on the Kaiser criterion, and then look at the loading matrix. It was revealed that there was only one independent factor for both set of features where PC1 pointed to trust, effective school leadership, and collaborative teachers for the school climate variables, and PC1 pointed to reading and math scores exceeded for achievement indicators. This could be summarised as a relationship between quality of teachers, and exceeding expectations in reading and math scores where the Pearson correlation between them is -0.367, implying that higher quality of teachers lead to lower test scores (?) Though, the data is represented in the PCA space, and not in the original coordinates.

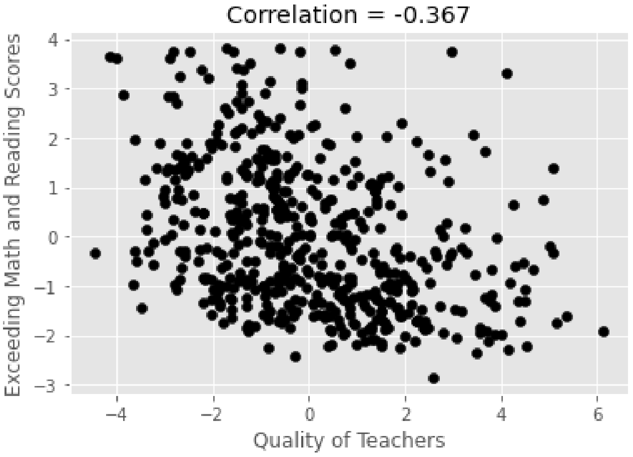


Fig 7: Impact of quality of teachers on exceeding reading and math scores in PCA space

By looking at the original data, and their correlation, it was revealed that all achievement indicators were positively correlated with school climate indicators, meaning that a higher average rating of “school climate” factors as perceived by the students will lead to a higher average student achievement on a state-wide standardized test, and a higher proportion of students exceeding state-wide expectations in reading and math. Though, the relationship might not be exactly linear since the magnitude of all correlation coefficients fall below 0.5. Most of the student achievement indicators seem slightly more correlated with having a supportive environment, and rigorous instruction than the rest of the school climate indicators.

	Rigorous Instruction	**Collab-Teachers**	**Supp - Environment**	**School Leadership**	**Strong family ties**	**Trust**
Student Achievement	0.390	0.262	0.484	0.187	0.202	0.240
Reading Scores	0.434	0.294	0.436	0.112	0.234	0.047
Math scores	0.412	0.272	0.416	0.090	0.211	0.048

Question 5:

Now we shall test our hypothesis that rich schools are more likely to get into HSPS since they spend more on students, and have access to better resources and quality teachers. Our null hypothesis assumes that there is no difference between rich and poor schools in being accepted to a HSPS whereas our alternative hypothesis assumes that rich schools have more acceptances compared to poor schools. In order to do this, we must create a categorical variable indicating 1 for a rich school and 0 for a poor school. To do this, per student spending is transformed into above and below the median, with the median spending per student (ignoring all nans) equal to $20147. An empty array was created for rich schools, and poor schools, where using a for loop, if it is a poor school (0) or rich school (1) append the number of acceptances to the array. Missing values were dropped across rows leading to even lengths for the arrays, as such an independent samples T-test was adopted since we had 2 independent groups. Using $\alpha = 0.05$ at a 95% significance level, the p-value was revealed to be 1.43e-1, with a T statistic of 6.925, meaning that we reject our null hypothesis of no difference, and the difference in outcomes is too large to be convincingly consistent with chance. The difference in sample means was (14.26), meaning that poor schools actually performed way better than rich schools, and this can also be seen in the graph below.

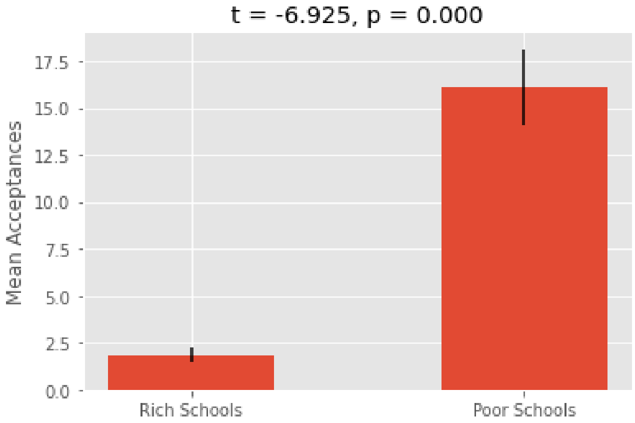


Fig 8: Difference in mean acceptances between rich and poor schools

Question 6:

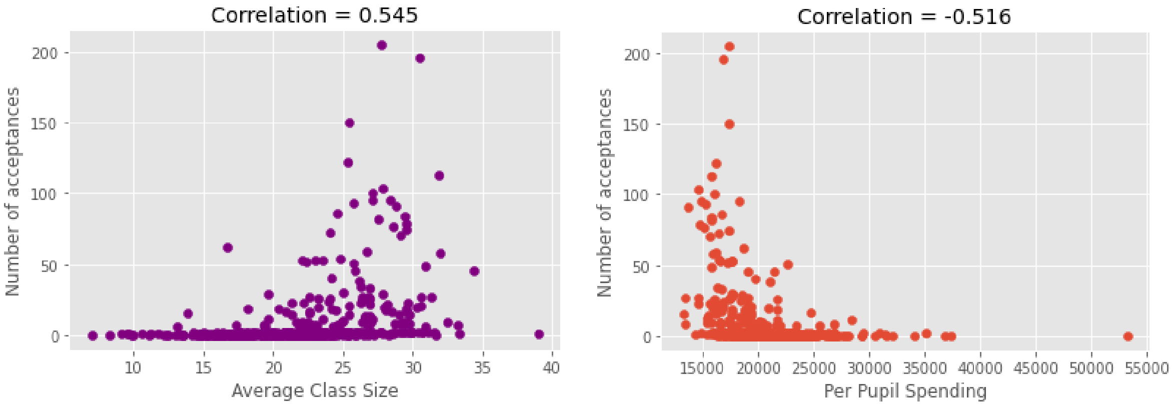


Fig 9 & 10: Impact of availability of resources on admission to HSPS

To find any any evidence that the availability of material resources, per pupil spending and class size impacts admission to HSPHS, the target variables were first added to a temporary array, and nans were dropped row-wise in order to plot them on a scatter plot to see any inherent structutre of the data. It seemed that the relationship was not exactly linear, as such, a spearman rho was suitable to show the relationship of the data where there was a positive non-linear relationship between average class size and admission to HSPS, and a negative non-linear relationship between per pupil spending and number of admissions to HSPS. In both graphs, there seems to be a cut off where after an average class size of 20, the number of acceptances to HSPS increases whereas after the median value of per pupil spending ($20147), the number of acceptances to HSPS seem to decrease.

Similarily, after reducing the dimensions of the achievement indicators using PCA, the loadings matrix revealed that PC1 pointed mostly towards exceeding reading and Math scores. After taking the old data in the new PCA space, a similar relationship as above was seen when it came to the relationship between availability of material resources and the achievement indicators. Computing the spearman rho between the two target variables, revealed a similar trend as above where per student spending negatively impacted achievement indicators, and average class size positvely impacted impacted achievement indicators.

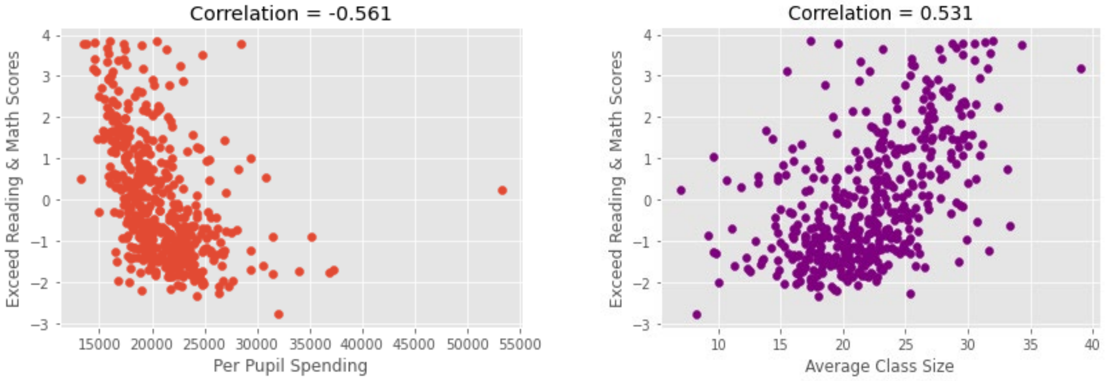


Fig 11 & 12: Impact of availability of resources on objective measures of achievement

Question 7:

It was revealed that 20.71% or 123 schools accounted for 90% of all students accepted to HSPS. To get this value, the dataframe values were assorted by number of acceptances in descending order first, then a column was added which consisted of the proportion accepted by each school based on the sum of all acceptances. Using a lazy counter, a for loop over the length of the sorted dataframe was implemented, with a conditon if the counter is <= 90, then append the desired values to a temporay dataframe, and increment the counter by the proportion accepted by the given school. Else, break out of the loop once the counter is equal to 90% of all students accepted to HSPS.

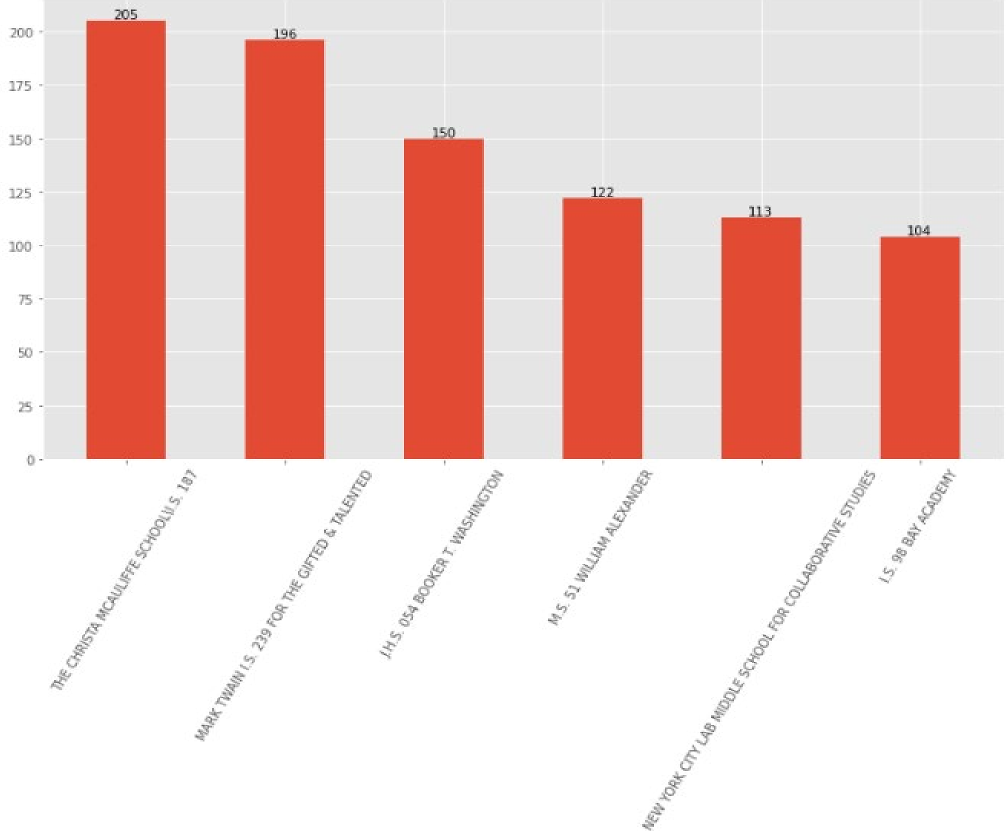
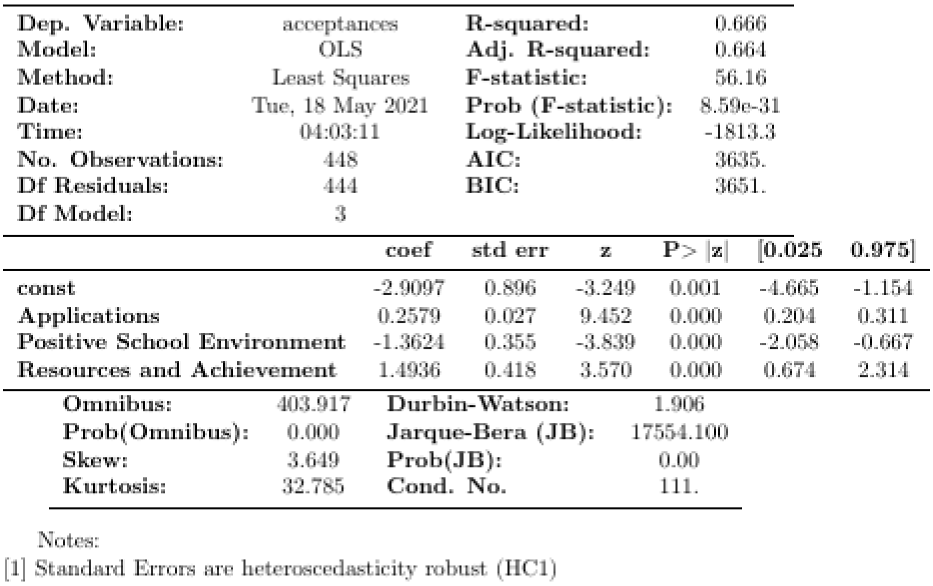


Fig 13: Top 6 schools in terms of number of acceptances to HSPS

Question 8:

Part A:

In building a model of our choice, we first have to make some assumptions, and take care of missing values row-wise to have equal number of rows across the different predictors. The model of choice is a multiple linear regression to predict the number of acceptances to HSPS by a school, where to deal with the inherent multicollinearity in the data, a PCA will be used to reduce the dimensions and find independent factors that explain most of the variance. For the purpose of this model, we will only look at factors such as the school climate variables, availability of material resources, achievement indicators and the number of applications. The demographic factors, school size, and other factors will not be used as they are not the main focus of this model. The dependent variable, the number of acceptances, was stored in an array along with the other independent variables. One of the predictors, the number of acceptances was used as it was, and the rest of the independent variables were reduced to fewer independent factors explaining most of the variance using a PCA. Based on the Kaiser criterion, only two principal components had eigenvalues greater than one. Upon investigating the loadings matrix, the first principal component could be interpreted as "Positive School Environment" whereas the second principal component pointed to several factors like availability of resources, trust, and exceeding math and reading scores. For the sake of completeness, the second PC could also be interpreted as "Resources and Achievement". After investigating the loadings matrix, the independent variables were created by using the original data in terms of the coordinate system spanned by the 2 principal components, and combining it with number of applications. Finally, using the sm method from statsmodels.api, the multiple linear regression was implemented.



```
Dep. Variable:          acceptances        R-squared:              0.666
Model:                  OLS                Adj. R-squared:         0.664
Method:                 Least Squares      F-statistic:            56.16
Date:                   Tue, 18 May 2021   Prob (F-statistic):     8.59e-31
Time:                   04:03:11           Log-Likelihood:         -1813.3
No. Observations:       448                AIC:                    3635.
Df Residuals:           444                BIC:                    3651.
Df Model:               3
                                coef     std err    z        P> |z|   [0.025    0.975]
const                           -2.9097  0.896      -3.249   0.001    -4.665    -1.154
Applications                    0.2579   0.027      9.452    0.000    0.204     0.311
Positive School Environment     -1.3624  0.355      -3.839   0.000    -2.058    -0.667
Resources and Achievement       1.4936   0.418      3.570    0.000    0.674     2.314
     Omnibus:            403.917   Durbin-Watson:        1.906
     Prob(Omnibus):      0.000     Jarque-Bera (JB):     17554.100
     Skew:               3.649     Prob(JB):             0.00
     Kurtosis:           32.785    Cond. No.             111.

   Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

Fig 14: Linear Regression table

Based on the table above, we can see that our model explained 66.67% of the variance in the dependent variable. Robust standard errors were used to correct for any inherent heteroskedasticity of the residuals. All the beta coefficients were statistically significant given that their p-values were lower than α = 0.05 at a 95% significance level. The coefficient on applications imply that, on average, an increase in the number of applications by one, is associated with a 0.2579 increase in the number of acceptances to HSPS. The coefficient on positive school environment imply that, on average, an increase in the number of applications by one, is associated with a 0.2579 increase in the number of acceptances to HSPS. The coefficient on applications imply that, on average, a more positive school environment, lowers the number of acceptances whereas the coefficient on resources and achievement imply that on average, more resources available per student, and a higher proportion of student achievement raises the number of acceptances to HSPS. The coefficients other than the applications are not directly interpretable given that the data is in PCA space, and not in the original space.

Part B:

A clustering model was used to determine which factors are most important in achieving high scores on objective measures of achievement. A temporary array of the original data was created where the missing values were dropped row-wise. An outcome variable named 'high performer' was created and coded 1 if the school had achievement indicators greater than the median values, and 0 otherwise. A predictor variable stored all the relevant predictors except the demographic, poverty, disability, and ESL variables to avoid complicating our model, and focus on availability of material resources, and school climate predictors instead. A PCA was conducted on the predictors to reduce the number of dimensions where there were only 2 principal components that met the Kaiser criterion. These components could be classified as overall teacher quality and availability of material resources. Using K-means clustering, and the silhouette scores, there were only two clusters that maximized the sum of silhouette scores. The clusters show that there may be two situations when it comes to determining how well a school scores on objective measures of achievement. Using a support vector machine, we can further classify the clusters where the data shows that there are more high performing schools that have a lesser overall school climate but have more availability of resources per pupil. Low performing schools have a somewhat better overall school climate but have few resources available per student. Though, there does not seem to be a clear decision boundary spanned by the support vectors as there is a lot of overlap in the data. The model was also 79% accurate at correctly classifying which school is a high or low performer based on school climate and availability of material resource, but this was an overestimation as it is not fitting on new data, thus it is fitting to noise, and cannot generalize.

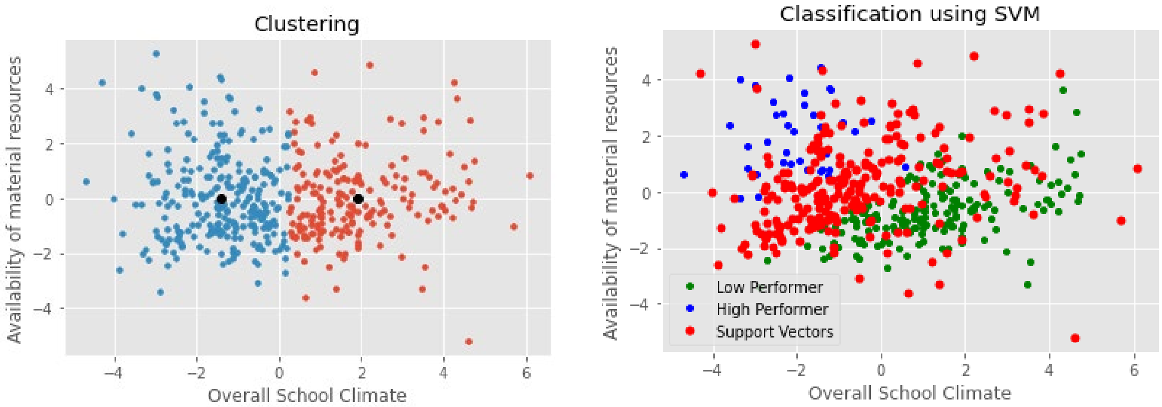


Fig 15 & 16: Clustering and Classification

Question 9:
Based on the answers thus far, it seems most reasonable that the first important characteristic that determines acceptance of a schools' students to a HSPS is the number of applications. On average, the more the students at a school actually apply to a HSPS, the higher the number of acceptances overall compared to schools who do not apply at all. For instance, if we take the example of the school with the highest per student odds, The Christa Mcauliffe School, it had 251 applications, 205 acceptances, and a school size of 873. Although, number of applications is certainly not the only important characteristic, and is also not the most significant one at improving the odds of acceptance. Based on question 6, a higher average class size seems to increase the number of acceptances to HSPS, and a higher per student spending seems to lower the number of acceptances. This is very counter intuitive as one would expect that if a school has a lower average class size and higher per student spending there would be more material resources at the disposal of students, giving them everything they need to improve the odds of acceptance, but this is not the case. Instead, having a positive school environment may compensate for this as having rigorous instruction, collaborative teachers, and supportive teachers creates an environment that encourages holistic learning, and also encourages students to apply these HSPSs. Finally, by combining better availability of material resources and a positive school environment, with a little bit of personal merit, students can also score better grades on standardised tests which can vastly improve the odds of acceptances to a HSPS.

Question 10:
As a data scientist for the New York City Department of Education, the first actionable recommendation to improve schools so that they send more students to HSPS is to create a more supportive, and collaborative environment where students holistically grow, and this in turn can help them score better grades on standardised tests and also encourage them to apply to a HSPS in the first place. These three things combined in an iterative procedure will allow schools to send more students to a HSPS. Increasing per pupil spending would not be the most efficient thing to do, as it seems like there is a negative relationship between spending and acceptances which may be due to the fact students are not taking advantage of their resources, or fail to realize the privilege they have compared to schools with lower per student spending who managed to have more acceptances to a HSPS.

The second actionable recommendation to improve on objective measures of achievement is again influenced by creating a positive, rigorous, and collaborative learning environment with a high quality of teachers that students can trust upon. It also seems that a higher average class size improves objective measures of achievements since there is more ground for collaboration, and student support. Likewise, based on question 8, improving the availability of material resources will also improve objective measures of achievement.